SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot

Year: 2023 | Citations: 984 | Authors: Elias Frantar, Dan Alistarh

---

Abstract

We show for the first time that large-scale generative pretrained transformer (GPT) family models can be pruned to at least 50% sparsity in one-shot, without any retraining, at minimal loss of accuracy. This is achieved via a new pruning method called SparseGPT, specifically designed to work efficiently and accurately on massive GPT-family models. We can execute SparseGPT on the largest available open-source models, OPT-175B and BLOOM-176B, in under 4.5 hours, and can reach 60% unstructured sparsity with negligible increase in perplexity: remarkably, more than 100 billion weights from these models can be ignored at inference time. SparseGPT generalizes to semi-structured (2:4 and 4:8) patterns, and is compatible with weight quantization approaches. The code is available at: https://github.com/IST-DASLab/sparsegpt.